Selection of KL neighbourhood in robust Bayesian inference

Submitted to Statistical Science

Selection of KL neighbourhood in robust Bayesian inference

Natalia A. Bochkina

The authors propose an attractive and coherent approach to robust inference in Bayesian statistics where potential joint misspecification in the likelihood and in the priors is reflected in misspecification of the posterior distribution and studied in its consequences in the decision problem of interest. I have a couple of comments, a short one (the first section) and a long one (the rest of the discussion).

1. DIRECTION OF ABSOLUTE CONTINUITY

Using one of the two possible Kullback-Leibler neighbourhoods determines whether the neighbourhood consists of the models that are absolutely continuous with respect to π_I, or whether π_I is absolutely continuous with respect to them; in the first case, with the choice $KL(\pi||\pi_I)$ mostly considered in the paper, the distributions in the KL neighbourhood of π_I must be absolutely continuous with respect to π_I. In fact, this is a consequence of the implicit assumption used in the proof of Theorem 4.2 which is the absolute continuity of one measure with respect to the other. Although it is not stated explicitly, it follows from the proof that the choice of the type of the continuity (of π with respect to π_I or of π_I with respect to π) results in the corresponding type of the KL divergence. An interesting question is whether there exists a unique divergence guaranteeing coherence if the assumption of absolute continuity of one measure with respect to the other is relaxed.

2. CHOICE OF THE SIZE OF THE KL NEIGHBOURHOOD

The second issue I would like to discuss is the choice of the size of the KL neighbourhood, C, of the distribution $\pi_I(\theta)$ over which the robustness of the decision is investigated. Due to duality, this problem is equivalent to the selection of λ_a in the least favourable distribution $\pi^{\sup}(\theta)$ (or in the corresponding "most favourable" distribution $\pi^{\inf}(\theta)$).

My main emphasis will be on the problem of misspecified likelihood $f(y \mid \theta)$ with the loss $L_y(\theta) = -\log f(y \mid \theta)$, as discussed in Section 4.1.3, although some methods apply to more general problems as considered by Watson and Holmes, with the least favourable distribution given by

$$\pi^{\sup}(\theta) \propto \pi(\theta)[f(y \mid \theta)]^{1-\lambda_a},$$

which is related to likelihood tempering. For a misspecified likelihood, this is a common way to compensate for model misspecification and to increase efficiency of inference about θ. The authors propose a way to choose λ_a in the context of

University of Edinburgh and Maxwell Institute, United Kingdom (e-mail: N.Bochkina@ed.ac.uk)

the DP approach to study the distribution of the loss which is equivalent to the choice of the total mass of the DP, by post-hoc diagnostic plots (Sections 4.3.2 and 5). I will discuss alternative ways of choosing λ_a.

There are two principally different cases of misspecified likelihood considered in the literature: the most natural situation when the true likelihood is unknown, and the less common case when the "true" parametric family is too complex to fit so a simplified model is used instead. These two cases are discussed below. For brevity, I will refer to the considered misspecified likelihood simply as likelihood.

2.1 True likelihood is unknown

2.1.1 Assumptions Calibration of inference about θ in the case of unknown true likelihood is widely considered in the literature, both frequentist and Bayesian (e.g. Royall & Tsou (2003); Mueller (2013)). The main aim for model calibration in such case is to achieve frequentist optimality, e.g. efficiency of parameter estimation.

Generally, in asymptotic framework for n iid observations, an inference under model misspecification is about the following value of the unknown model parameter:

$$\theta_0 = \arg\min_{\theta} \lim_{n\to\infty} KL(P^n_{\text{true}}||P^n_{\theta}).$$

A nonasymptotic and non-iid version can be found in Panov & Spokoiny (2015). Generally, θ_0 may be different from the "true" value of the parameter θ_{true} of the correctly specified model P_{true}. One of the main assumptions of the calibration methods is that the inference about θ_0 is meaningful. This assumption is satisfied, for instance, for models with a location parameter θ and a symmetric distribution with $\theta_0 = \theta_{true}$. Therefore, the main issue is adjustment of variability that characterises uncertainty of inference on θ under the misspecified model where the variability is often underestimated, leading to overconfident decisions.

In addition, an implicit assumption in these papers is regularity of both true and misspecified models, namely that $\nabla_\theta \log f(y \mid \theta_0)$ has zero expected value and finite variance.

2.1.2 One dimensional parameter For one-dimensional parameter θ, the problem of calibrating a regular likelihood (mostly from the frequentist perspective) was considered by Royall & Tsou (2003) who showed that the inference based on the likelihood tempered by the temperature H/V with

$$\begin{aligned} H &= -\mathbb{E}_{P_{\text{true}}} \nabla^2_\theta \log f(y \mid \theta_0) = \mathbb{E}_{P_{\text{true}}} \nabla^2_\theta L_y(\theta_0), \\ V &= \mathbb{E}_{P_{\text{true}}} [\nabla_\theta \log f(y \mid \theta_0)]^2 = \mathbb{E}_{P_{\text{true}}} [\nabla_\theta L_y(\theta_0)]^2 \end{aligned} \tag{1}$$

is asymptotically efficient. In the case of iid observations unknown H and V can be consistently estimated by $\hat{H} = n^{-1} \sum_i \nabla^2_\theta L_{y_i}(\widehat{\theta})$ and $\hat{V} = n^{-1} \sum_i [\nabla_\theta L_{y_i}(\widehat{\theta})]^2$ where $\widehat{\theta}$ is a consistent estimate of θ_0, e.g. the quasi-MLE. The authors also suggest that thus adjusted likelihood should be used as the likelihood in Bayesian inference.

In the notation of the discussed paper, this corresponds to the choice $\lambda_a = 1 - H/V$. Often, $H/V < 1$, i.e. the misspecified model contains less information than the correct model, and hence the corresponding posterior distribution corresponds to the least favourable prior. One interesting question is whether there exist cases with $H/V > 1$ which would correspond to the maximin distribution $\pi^{\text{inf}}(\theta)$.

2.1.3 Multivariate parameter Tempering of the misspecified likelihood in the case of multivariate θ for composite likelihoods was considered by Ribatet et al (2012) who referred to it as the magnitude adjustment. A more precise calibrating procedure of the unknown likelihood parametrised by a multivariate θ was considered e.g. by Mueller (2013) for regular models where the distribution of a sufficient statistic for θ, the quasi-MLE, is asymptotically Gaussian and effectively is used as the "adjusted" likelihood. The proposed adjustment method is the linear change of variables $\pi_{adj}(\theta \mid y) = \pi_I(A\theta \mid y)$ where in the case of (asymptotically) non-informative prior, A is such that the variance of the adjusted posterior is equal to the sandwich covariance matrix. This corresponds to the smallest frequentist risk and is the variance of the quasi-MLE, i.e. $A^T H A = V$ where matrices H and V are defined by (1). In case of an informative prior, denoting $B = \nabla^2 \log \pi(\theta_0)$, the matrix A should satisfy $A^T[H + B]A = V + B$. In practice, H (or $H + B$) can be estimated by the posterior covariance matrix, B can be estimated by $\nabla^2 \log \pi(\hat{\theta})$ where $\hat{\theta}$ is a consistent estimator of θ_0, e.g. the posterior mean, and for iid observations V can be estimated by $n^{-1} \sum_i [\nabla L_{y_i}(\hat{\theta})]^2$.

The corresponding magnitude adjustment for tempered likelihood is $1 - \lambda_a = ||HV^{-1}||$ for the asymptotically noninformative prior, and $1 - \lambda_a = ||(H+B)(V+B)^{-1}||$ for tempered posterior with an informative prior which can be estimated as suggested above. Asymptotically, this can also be viewed as the approach proposed in the discussed paper with $a = \hat{\theta}$ and $L_a(\theta) = ||\theta - \hat{\theta}||_Q^2 = (\theta - \hat{\theta})^T Q (\theta - \hat{\theta})$ with $Q = H$ or $Q = H + B$, respectively. A different Q would result in a different value of λ_a. An interesting question, which also applies to some examples below, is whether it is possible to apply or to extend the discussed approach to accommodate the more precise adjustment of the misspecified posterior.

Viele (2007) proposed a method to fit a posterior distribution of KL divergence $KL(P_{\text{true}}||P_\theta)$ given a sample from P_{true} which can be applied as an alternative way of selecting the upper bound on the KL neighbourhood size, C (and hence on $\lambda(C)$), for instance by taking C as 95% percentile of the posterior distribution of the KL distance. This method is developed for a Dirichlet process prior for P. Hence, it should be possible to adjust the approach of Watson & Holmes to calibrating $\pi_I(\theta)$ based on the DP prior (Section 4.3).

2.1.4 PAC-Bayesian approach and theoretical bound on λ_a In the absence of the likelihood, the maximin distribution $\pi^{inf}(\theta)$ corresponds to PAC-Bayesian estimation and Gibbs posteriors (Section 4.2.2). The robust perspective given in the discussed paper gives an insight into this approach which corresponds to the most optimistic posterior distribution, with the smallest possible value of the posterior loss in the given neighbourhood of the posterior. The discussed approach would naturally suggest using also the corresponding least favourable distribution however in many cases it may not be proper. A value of λ_a can be chosen by adapting the model calibration approaches discussed above applied to $\exp(-L_a(\theta))$ as the misspecified likelihood.

For some PAC-Bayesian estimators, conditions on λ_a are available that result in optimal inference, from the frequentist perspective, about the unknown parameter with respect to the loss function $L_a(\theta)$. For instance, for sparse high dimensional linear regression and ℓ^2 loss, theoretical upper bounds on λ_a under general conditions on the true likelihood can be derived from Dalalyan & Tsybakov (2010).

2.2 Known true model

The less common case is when the true model is known, for instance, in approximate computation, with the typical case being composite likelihood models. Ribatet et al (2012) proposed the magnitude adjustment which corresponds to tempering of the posterior with $1-\lambda_a = ||\Omega_{\text{true}}\Omega_{miss}^{-1}||$ where Ω_{miss} and Ω_{true} are posterior precision matrices of the approximate and the true posterior respectively. Stoehr & Friel (2015) proposed an affine adjustment so that the first two moments of the posterior based on the approximate model match the corresponding moments of the true posterior distribution.

The calibration methods above are based on local asymptotic normality of the posterior distribution. Similar calibration is possible for posteriors concentrating around $\widehat{\theta}$ in a different way. A typical example is an asymptotically exponential or gamma posterior distribution:

$$(2)\ \pi_I(\theta) = \prod_{j=1}^{p} v_j^{\alpha_j - 1} \exp\{-\sum_{j=1}^{p} \kappa_j v_j + O_P(1)\} I(v_j \geq 0 \text{ for } j = 1, \ldots, p),$$

where $v = \theta - \theta_0 - \Delta_y$, usually $\kappa_j = -\mathbb{E}_{P_{\text{true}}} \nabla_{\theta_j} \log f(y \mid \theta_0)$ and $\alpha_j > 0$ are determined by the prior. For instance, for a locally asymptotically exponential (LAE) likelihood for parameter θ, e.g. for a likelihood with a jump at θ, and a locally constant prior, the posterior has form (2) with all $\alpha_j = 1$ and random Δ_y (Chernozhukov & Hong, 2004). This also holds when θ_0 is a sharp maximum of $\mathbb{E} \log f(Y_i \mid \theta)$ which is usually attained on the boundary of the parameter set. For instance, this happens for a truncated Gaussian likelihood when the "true" value of the parameter is outside of the truncation, or when the observations are independent Poisson$(A_i\theta)$ random variables such that $A_i\theta_0 = 0$ for some i; in this case, the posterior is of the form (2) with $\Delta_y \equiv 0$ and $\alpha_j > 0$ can be arbitrary (Bochkina & Green, 2014). If the posterior corresponding to the true likelihood is of the form (2) with $p = 1$ and $\kappa_1 = \kappa_*$, the misspecified likelihood should be tempered by $1 - \lambda_a = \kappa_*/\kappa_0$ which for a known true likelihood and one-dimensional θ can be estimated by $\widehat{\kappa_*/\kappa_0} = \frac{\nabla_\theta \log f_{\text{true}}(y|\hat{\theta})}{\nabla_\theta \log f(y|\hat{\theta})}$ with $\hat{\theta}$ being a consistent estimator of θ_0, e.g. the quasi-MLE. For a LAE likelihood, the random bias Δ_y is assumed to be the same or to have the same distribution as under the correctly specified model. A similar adjustment is possible in a multivariate case.

A related problem is calibration of a known desirable prior, $\pi_I(\theta)$, to fit prior expert information of the form $\mathbb{E}_\pi g(\theta) = 0$ where the expectation is taken with respect to a prior. Choi (2015) proposes to find a prior that satisfies such constraints and which is closest to the desirable prior in KL distance. The function g can depend on the likelihood family (but not on the observed data). This optimisation problem is dual to the problem considered in the discussed paper, with the unconstrained Lagrangian dual of the similar form $\pi = \arg\inf_\pi [KL(\pi||\pi_I) + \eta g(\theta)]$, where $g(\theta) = L_a(\theta) - \mathbb{E}_\pi L_a(\theta)$, with the optimal prior $\pi(\theta) \propto \pi_I(\theta) \exp(\lambda_*^T g(\theta))$ and $\lambda_* = \arg\min_\lambda \mathbb{E}_{\pi_I} \exp\{\lambda^T g(\theta)\}$. The author proposes the following estimator of λ_* based on N Monte Carlo draws from π_I:

$$\hat{\lambda} = \arg\min_\lambda N^{-1} \sum_{j=1}^{N} \exp\{\lambda^T g(\theta_j)\}.$$

I think that the authors have brought to discussion an interesting topic of robustness that is not routinely addressed by Bayesian statisticians in this form; it is more general that the usual checks to the sensitivity of the prior by simulations, usually over a (relatively) small number of possible alternative scenarios. I hope this will motivate further methodological development and routine reports of sensitivity of the decision making to model misspecification in practice.

REFERENCES

Bochkina, N.A. and Green, P.J. (2014). *The Bernstein - von Mises theorem and nonregular models*. Ann.Statist. 42:5, 1850-1878

Chernozhukov, V. and Hong, (2004). *Likelihood Estimation and Inference in a Class of Nonregular Econometric Models*. Econometrica 72:5, 1445-1480

Choi, Hwan-sik (2015). *Expert Information and Nonparametric Bayesian Inference of Rare Events*. Bayesian Analysis Advance Publication, 26 May 2015.

Dalalyan, A. and Tsybakov, A.B. (2010). *Sparse regression learning by aggregation and Langevin Monte-Carlo*. J. Comput. System Sci 78:5, 1423-1443

Mueller, U. (2013) *Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix*. Econometrica 81:5, 1805-1849

Panov, M. & Spokoiny, V. (2015) *Finite Sample Bernstein von Mises Theorem for Semiparametric Problems*. Bayesian Analysis 10:3, 665-710.

Ribatet, M. and Cooley, D. and Davison, A.C.(2012) *Bayesian inference from composite likelihoods, with an application to spatial extremes*. Statistica Sinica 22, 813-845

Royall, R.M. and Tsou, T-S(2003) *Interpreting statistical evidence using imperfect models: Robust adjusted likelihood functions*. JRSS-B 65, 391-404

Stoehr, J. and Friel, N.(2015) *Calibration of conditional composite likelihood for Bayesian inference on Gibbs random fields*. AISTATS, JMLR: W&CP, 921-929

Viele, K. (2007) *Nonparametric estimation of Kullback-Leibler information illustrated by evaluating goodness of fit*. Bayesian Analysis 2:2, 239-280.